News/Lab Note for inclusion on MIT Lincoln Laboratory's external website
26 April 2016

Finding relevant data in a sea of languages

A cross-language search engine combines language identification, machine translation, information retrieval, and query-biased summarization techniques to enable English monolingual analysts to find foreign language documents relevant to their investigations.[1]

"About 6,000 languages are currently spoken in the world today," says Elizabeth Salesky of Lincoln Laboratory's Human Language Technology (HLT) Group. "Within the law enforcement community, there are not enough multilingual analysts who possess the necessary level of proficiency to understand and analyze content across these languages," she continues.

This problem of too many languages and too few specialized analysts is one Salesky and her colleagues are now working to solve for law enforcement agencies, but their work has potential application for the Department of Defense and Intelligence Community. The research team is taking advantage of major advances in language recognition, speaker recognition, speech recognition, machine translation, and information retrieval to automate language processing tasks so that the limited number of linguists available for analyzing text and spoken foreign languages can be used more efficiently. "With HLT, an equivalent of 20 times more foreign language analysts are at your disposal," says Salesky.

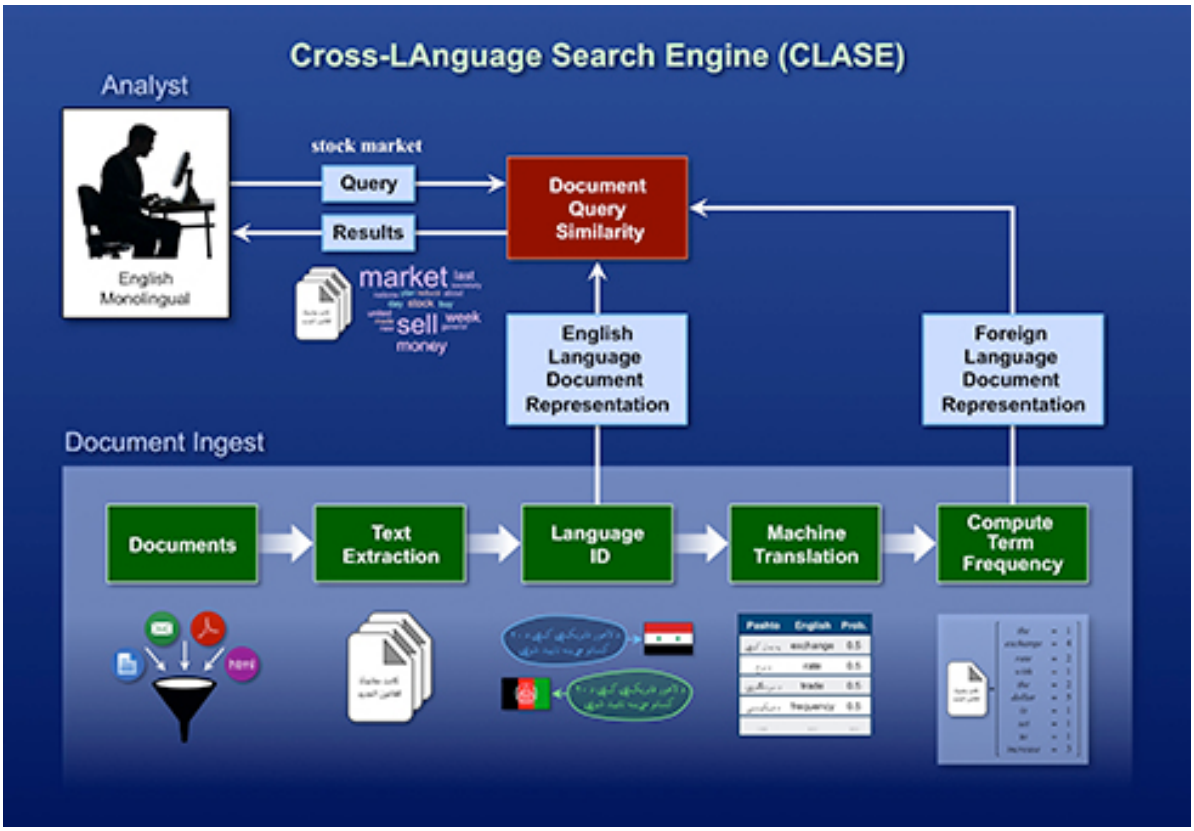
One area in which Laboratory researchers are focusing their efforts is cross-language information retrieval (CLIR). The Cross-LAnguage Search Engine, or CLASE, is a CLIR tool developed by the HLT Group for the Federal Bureau of Investigation (FBI). CLASE is a fusion of Laboratory research in language identification, machine translation, information retrieval, and query-biased summarization. CLASE enables English monolingual analysts to help search for and filter foreign language documents—tasks that have traditionally been restricted to foreign language analysts.

Laboratory researchers considered three algorithmic approaches to CLIR that have emerged in the HLT research community: query translation, document translation, and probabilistic CLIR. In query translation, an English-speaking analyst queries foreign language documents for an English phrase; that query is translated into a foreign language via machine translation. The most relevant foreign language documents containing the translated query are then translated into English and returned to the analyst. In document translation, foreign language documents are translated into English; an analyst then queries the translated documents for an English phrase, and the most relevant documents are returned to the analyst. Probabilistic CLIR, the approach that researchers within the HLT Group are taking, is based on machine translation lattices (graphs in which edges connect related translations).

First, foreign language documents are translated into English via machine translation. The machine translation model projects foreign words into English probabilistically and then outputs a translation lattice containing all possible translations with their respective probabilities of accuracy. "For example, the lattice for the French word *capacité* would show connections to and probability scores for the English words *capacity* and *ability*," says Michael Coury of the HLT Group. On the basis of an analyst's query of a document collection, the documents containing the most probable translations would be

[1] This work was sponsored by the Department of Justice under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

extracted from the collection for analysis, even if they contain the second or third most likely translation candidates. This method allows analysts to retrieve documents not found by query or document translation. CLIR results are evaluated on the basis of precision (the fraction of retrieved documents that are relevant), recall (the fraction of relevant documents that are retrieved), and F-measure (the harmonic mean of precision and recall).



An English-speaking FBI analyst searches through a large collection of potential "white collar" crime evidence that has been reported in documents written in many different foreign languages. The analyst enters a query for a term related to the crime (stock market). The Cross-LAnguage Search Engine (CLASE) has already preprocessed the documents, extracting text to identify the language in which they were composed, translating the documents to English (if not originally in English), and then indexing them by recurring relevant terms (this sequence is depicted in green in the lower box). CLASE then supplies the analyst with documents (either ones in English or foreign language ones represented in English) that contain exact matches to or words related to the query term.

"We are interested in achieving high recall. If we do not retrieve all relevant documents, we could miss a key piece of evidence," says Coury. "When we search on Google, we are usually only interested in the 10 most relevant results on the first page. For the law enforcement community, we want to identify every single potentially meaningful search result."

As mentioned previously, CLASE is heavily dependent upon the Laboratory's research in language identification and machine translation. Jennifer Williams, also in the HLT Group, has been developing algorithms to identify the languages present in text data so that the appropriate machine translation

models can be selected by CLASE. According to Williams, text language identification faces many challenges. Reliable methods are needed for improving the accuracy of distinguishing between languages with similar character sets. Differentiating between similar languages is not the only problem for text language identification. Another challenge involves processing user-generated content that has been Romanized, or transcribed into the Latin alphabet, on the basis of phonetics. "One example of this practice is tweets written in Romanized Arabic, referred to as 'Arabizi' in the HLT community. We see Romanization with Chinese, Russian, and other languages as well," says Williams. In some cases, ground truth data on languages is nonexistent (e.g., for low-resource languages, such as Urdu and Hausa) or is unreliable. "No universal language identification system exists, so the variances between different systems can be extreme," she adds.

Other researchers in the group are creating systems to automatically translate text from one language to another. According to Salesky, these efforts in machine translation have been critical to the HLT Group's work in CLIR. Wade Shen, an associate leader of the HLT Group who is currently serving an Intergovernmental Personnel Act assignment at the Defense Advanced Research Projects Agency, and university researchers have developed an open-source statistical machine translation toolkit called Moses. This phrase-based system allows users to train translation models for any language pair and find the highest-probability translation among the possible choices.

A problem inherent to training translation models for the FBI is the mismatch between the domain from which available training data are drawn and the domain in which the FBI is interested. A domain in this context refers to a topic or field that has its own writing style, content, and conventions. For example, tweets are limited to 140 characters and are written in a casual style that often contain abbreviations and misspellings; news articles are fairly long and lead with important information; and police reports are composed in a formal style and contain unique terminology. According to Jennifer Drexler, a member of the HLT Group who is pursuing an advanced degree at MIT under the Lincoln Scholars program, translation accuracy is best when the domain from which training data are acquired is similar to the domain in which the data of interest reside. Such a matchup helps to create a translation model that is informed about the nuances and peculiarities within the target domain. However, acquiring training data in the domain of interest can be difficult and expensive. It takes millions of parallel human-translated documents to create an automatic translation model. Human translation can cost between $0.20 and $0.80 per word. For rare languages, such as Urdu, translation costs are at the high rate to reward translators for their specialized knowledge.

Drexler and Shen, in collaboration with government researchers, found that hierarchical maximum a posteriori (MAP) adaptation[2] could be used to improve translation results when the amount of training data in the domain of interest is limited, but large amounts of data from other domains are available. This is exactly the case for the CLASE system—there are relatively small amounts of "in-domain" FBI data that can be used to train a translation model because of the security considerations that limit translators' access to in-domain data, but "out-of-domain" data (e.g., news articles or blogs) are much more abundant. The hierarchical MAP adaptation technique provides a principled way of combining models from these different domains, such that the final model is biased towards using the in-domain data whenever possible but is able to take advantage of the out-of-domain data when necessary.

[2] A full discussion of the hierarchical MAP adaptation algorithm and results of the experiments carried out to evaluate the algorithm can be found in A.R. Aminzadeh, J. Drexler, T. Anderson, and W. Shen, "Improved Phrase Translation Modeling Using MAP Adaptation," *Text, Speech and Dialogue*, pp. 394-402. Berlin Heidelberg: Springer, 2012.

Shen and former Lincoln Laboratory staff member Sharon Tam began the HLT Group's work in CLIR during the early 2010s. Researchers in the HLT community had previously shown document translation to be more accurate than query translation; therefore, Shen and Tam focused on evaluating how document translation compared to probabilistic CLIR. They found that probabilistic CLIR offered at least a 30% improvement in precision as compared to document translation, so they made the decision to use the probabilistic CLIR algorithm for CLASE.

Since joining Lincoln Laboratory in 2012, Coury has built upon Shen and Tam's initial experiments to evaluate CLIR performance on documents pertaining to an FBI case. The results are classified, but the HLT Group is confident that their CLIR technique is state of the art and that CLASE is a valuable tool for FBI analysts to use during document triage. "Our probabilistic approach was shown to be critical to retrieving documents cross language. For the very first time, FBI monolinguals can assist in document triage, adding a much larger pool of analysts to the smaller body of foreign language specialists," says Coury.

CLIR research has led to the related problem of how to present retrieved content to an analyst—a problem that Williams, Shen, and Tam began researching in 2013. Williams continues leading this effort to define the relationship between query-biased summarization and overall system performance as a human-in-the-loop problem. Williams and colleagues found query-biased summarization algorithms can be used to automatically capture relevant content from a document when given the analyst's query and to then present that content as a condensed version of the original document. "Search engines use this kind of summarization, providing snippets with links to the websites containing your search terms," says Williams.

To evaluate the utility of query-biased summaries for CLIR, the team ran experiments to compare 13 summarization methods falling under the following categories: unbiased full machine-translated text, unbiased word clouds, query-biased word clouds, and query-biased sentence summaries. They discovered query-biased word clouds to be the best overall summarization strategy in terms of recall, time on task, and accuracy.[3] However, users have different preferences or needs when it comes to digesting information, as evidenced by Williams herself, who does not like word clouds. Some users may prefer sentences while others may prefer an auditory signal rather than a textual or visual representation of information.

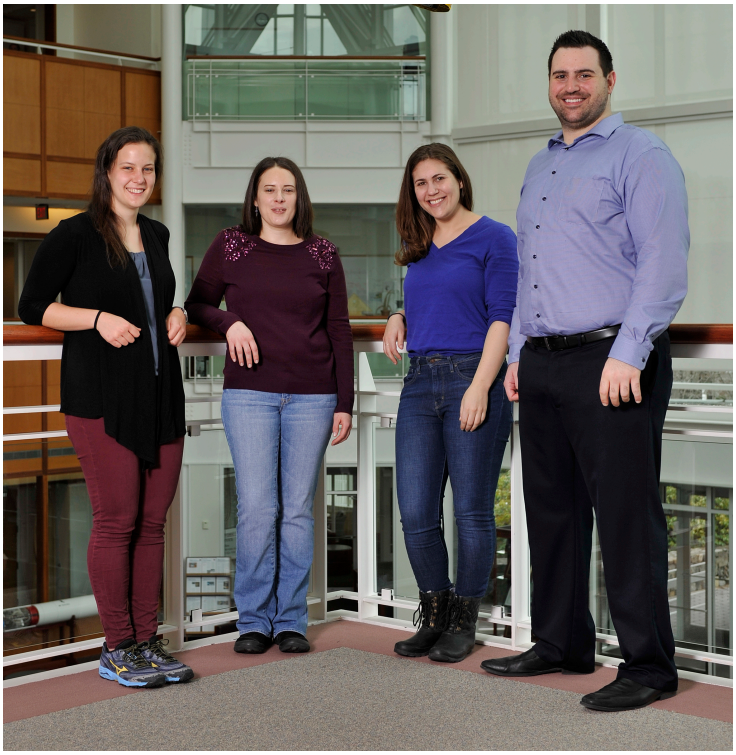
"Cross-language query-biased summarization is an important part of CLASE because it helps analysts decide which foreign language documents they should read. We can leverage this summarization to improve overall system recall," explains Williams. While in theory query-biased summarization could enable an analyst to work faster, additional research is required to determine if such summarization is practical for real-world CLIR systems such as CLASE.

According to Coury, there are many real-world scenarios that could benefit from using CLASE. "You could imagine it being used during the Syrian refugee crisis. Keyword searches could be performed on collected Twitter feeds to help analysts find potential terrorists hiding among migrant groups," he says. Coury and his colleagues are also interested in how the technology could benefit humanitarian

[3] A full discussion of the results of the experiments carried out to evaluate summarization methods can be found in J. Williams, S. Tam, and W. Shen, "Finding Good Enough: A Task-Based Evaluation of Query Biased Summarization for Cross Language Information Retrieval," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Oct. 2014.

assistance and disaster relief efforts—quickly retrieving information during crises involving languages for which translators are scarce and no automated HLT technology exists.

As Laboratory researchers continue to make advances in machine translation, CLIR, and query-biased summarization, these advances will be incorporated into CLASE and will continue to help analysts quickly and accurately find the information they need. “I noticed when I was searching through the HLT literature that a research team would do a study and stop short,” says Williams. “Each study was trying to solve a very specific problem. No single work combined machine translation, information retrieval, and query-biased summarization. Lincoln Laboratory is the first to draw all of these areas together.”



The Cross-Language Search Engine (CLASE) team includes, left to right, Elizabeth Salesky, Jennifer Williams, Jennifer Drexler, Michael Coury.